Use of Kalman Filtering and Particle Filtering in a Benzene Leachate Transport Model

Shoou-Yuh Chang [1], and Sikdar Latif[*2]

[1]Dept. of Civil Engineering, North Carolina A&T State University, Professor

[*2]Dept. of Civil Engineering, North Carolina A&T State University, Research Assistant

[1]1601 E Market St., 438 McNair Hall, NC 27411

[1]chang@ncat.edu; [*2]rajib_901@yahoo.com

Abstract

Groundwater contamination is one of the major environmental risks related to landfills. Basic information about the behavior of pollutants in soil-groundwater is needed in order to evaluate the migration of leachate from landfills and to establish efficient groundwater monitoring systems. However, in practice, it is very difficult to get exact subsurface data. Thus, modeling the behavior of pollutants during the flow of leachate through soil is important in predicting the fate of the pollutants. In this study, a one-dimensional transport model with advection and dispersion was used as the deterministic model of benzene leachate transport from an industrial landfill. A particle filtering (PF) with sequential importance resampling (SIR) filter and discrete Kalman filtering (KF) were proposed to improve the prediction of the benzene plume transport. A traditional root mean square error (RMSE) of benzene concentration is used to compare the effectiveness of the KF, PF, and a conventional numerical model. The results showed that Kalman filtering outperformed Particle filtering in the initial time steps. Both KF and PF can reduce the error up to 80% in comparison to a conventional numerical approach.

Keywords

Leachate Transport; Groundwater; Kalman Filtering; Particle Filtering

Introduction

Landfills that receive municipal andindustrial waste, are one of the most severe sources of ground water pollution due to leachate, the liquid produced when water percolates through solid waste landfills. Some of the more common soil contaminants are chlorinated hydrocarbons (CFH), heavy metals, Methyl Tertiary Butyl Ether (MTBE), zinc, arsenic, and benzene. Water and soil contamination are important pathways of concern for transmission of benzene. In the U.S. alone, there are approximately 100,000 different sites with benzene soil or groundwater contamination (Benzene: Encyclopedia 2009). Subsurface contaminant transport models play an important role in explaining how a contamination plume evolves, by evaluating the likely behavior of systems for contaminant remediation, predicting how the contaminant will behave in the future, and assessing the risk of the contaminant accuratly.

Mathematical deterministic models are widely used in the subsurface contaminant transport process. The predictions by these models may deviate from the real field value. These unavoidable prediction errors may arise from inaccurate assumption of different criteria and parameters, space and time limits of numerical schemes, and boundary conditions. With the system model alone, it is very difficult to predict the true dynamic state of the pollutant. Therefore, observational data is needed to guide the deterministic system model to assimilate the true state of the contaminant. The available databases are growing, but typically incomplete and contain measurement uncertainties. Use of these databases for verification of dynamic, multivariable models is difficult with the traditional qualitative and deterministic models.

Filtering is the most applicable to deal with a low dimensional system with a well known model and a dense data base, for which highly accurate short-term forecasts are required (Schrader and Moore, 1977). Although conventional numerical models and estimation techniques may sometimes provide good solutions for many water quality problems, filtering techniques combined with traditional numerical approaches can provide cost-effective solutions where the construction of observation wells is very costly and sometimes impossible. Contaminant state estimation can be considered as an optimal filtering problem within a Bayesian framework. The most well-known Bayesian state-estimation algorithm is the KF which is very effective for linear systems with Gaussian distribution. PF is generally applied for nonlinear and non Gaussian cases. Since several unknown and random hydro-geological factors and parameters are

associated with subsurface contaminant migration, the system behavior cannot be categorized into a specific linear or nonlinear system. Therefore, both filtering techniques in conjunction with a traditional deterministic numerical model are examined in this study.

Both KF and PF are generally applied in signal processing and engineering, as well as in biology, biochemistry, structure modeling, geosciences, immunology, materials science, chemical process modeling, pharmacology, toxicology, and social science. For the last three decades, KF and PF have been applied in surface and subsurface hydrologic systems and water quality modeling (Van Geer, 1982; Cosby et al., 1984; Whitehead and Hornberger, 1984; Yu et al., 1989; Graham and McLaughlin, 1989; Yangxiao et al., 1991; Zou and Parr, 1995; Ferraresi and Marinelli, 1996; Harrouni et al., 1997; Porter et al., 2000;Walker et al., 2001; McLaughlin, 2002; Cheng, 2002; Chang and Jin, 2005; Rozos and Koutsoyiannis, 2011; Li et al., 2012; Panzeri et al., 2013).

The objectives of this study are to examine the effectiveness of KF and PF in subsurface leachate transport modeling and to compare the performance of KF, PF, and a conventional numerical model in a one dimensional leachate model.

Methodology

The one-dimensional form of the advection-dispersion equation for benzene leaching in saturated, homogeneous, isotropic materials in uniform flow as shown in 1 is described by the following partial differential equation (PDE):

$$\frac{\partial C}{\partial t}=\frac{D_y}{R}\frac{\partial^2 C}{\partial y^2}-\frac{V}{R}\frac{\partial C}{\partial y} \tag{1}$$

where C = solute concentration, mg/L; t = time, d; y = cartesian coordinate direction along the flow line, m; D_y = dispersion coefficients in y direction, m²/d; V = linear velocity of flow field in the y direction, m/d; R = retardation factor, dimensionless.

The boundary conditions of the one-dimensional mass transport equation with an instantaneous point source are expressed as (at t = 0) (Schwartz and Zhang 1994): $\frac{\partial C}{\partial y}=0$, at $y = 0$ and ∞ with the initial condition (at t = 0):

$$M= M_0,\ C=C_0, \text{ at } y = 0;$$

$$C=0, \text{ at } 0 < y <\infty;$$

The analytical solution for the governing partial differential equation (PDE) is given by (Schwartz and Zhang, 1994):

$$C(y,t)=\frac{M_0}{\eta A\sqrt{4\pi D_y tR}}e^{-\frac{(y-vt/R)^2}{4D_y t/R}} \tag{2}$$

where η is the porosity of soil medium, and A is the cross-sectional area of the one-dimensional model.

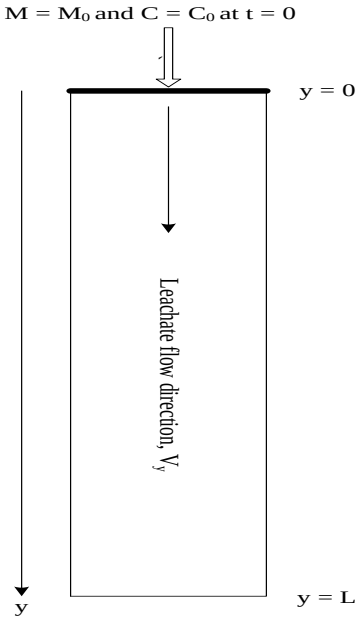

FIG. 1 MASS TRANSPORT FROM A POINT SOURCE

Solving the one-dimensional subsurface transport equation numerically by forward time centered-space (FTCS) difference method; the following solution can be used,

$$\mathbf{C_{i,t+1}}=\lambda_1\mathbf{C_{i-1,t}}+\lambda_2\mathbf{C_{i,t}}+\lambda_3\mathbf{C_{i+1,t}} \tag{3}$$

where $\mathbf{C_{i,t}}$= vector of pollutant concentration at node i and time t; $\mathbf{C_{i,t+1}}$= vector of pollutant concentration at node i and time $t+1$; $\mathbf{C_{i+1}}$= vector of pollutant concentration at node $i+1$ and time t;

$$\lambda_1=\frac{\Delta t D_y}{\Delta y^2 R}+\frac{\Delta t V}{2\Delta y R};\ \lambda_2=1-\frac{2\Delta t D_y}{\Delta y^2 R};$$

$$\lambda_3=\frac{\Delta t D_y}{\Delta y^2 R}-\frac{\Delta t V}{2\Delta y R};$$

i= space co-ordinate of node; t = time step co-ordinate.

Equation (3) can be rewritten in the following state-space form:

$$\mathbf{x_{t+1}}=\mathbf{A_t x_t} \tag{4}$$

where $\mathbf{x_{t+1}}$ and $\mathbf{x_t}$ are the state variables described as vectors of contaminant concentration at all nodes in the problem domain at time $(t+1)\Delta t$ and $t\Delta t$, respectively; $\mathbf{A_t}$ is the state transition matrix which gives a finite difference scheme to move a time step to the next.

One-dimensional leachate contaminant plumes can be

simulated by assuming that the system is stochastic. Stochastic processes are used for modeling complicated real-world transport processes with uncertain sources of errors. In order to analyze and make inference about a dynamic system, at least two models are required; a model describing the evolution of the state with time (the system model) and, a model relating the noisy measurements to the state (the measurement model) needed to analyze a dynamic system (Arulampalam et al., 2002).

The above advection-dispersion model of contaminant transport is a function of the retardation factor, dispersion and velocity. In the PDE model, it is assumed that the velocity remains constant in the y direction, the retardation factor (*R)* is constant, the soil layer is homogenous, and isothermal condition prevails. Furthermore, the soil layer is assumed to be completely saturated and benzene is assumed to be a non-reactive contaminant. However, in the field none of those incidences can take place. Again the boundary conditions may not prevail throughout the transport process. Subsurface consists of heterogeneous layers, each exhibiting different properties. Contaminant transport modeling cannot account for all of those phenomena. This is why the transport process should be simulated as a stochastic system. Our idealistic model contains error with respect to the previously mentioned assumptions which conflict with the actual true field value. Again the partial differential equation (PDE) is solely the function of a few parameters like advection, dispersion and retardation. Many other parameters and factors, i.e., degradation, adsorption, volatility of solute, soil texture and so on, affect contaminant transport in the field. Any uncertainty analysis deals with random variables. The specific values of the random variables cannot be obtained. Only the statistical nature of the random system can be found. In our model, a standard procedure of signal processing (dynamic state estimation) is followed (Welch and Bishop, 1995). Although errors vary spatially and temporally, there is no definite mathematical function to relate the errors with space and time. Therefore, a probabilistic approach is followed to take those into account.

To simulate our process or prediction equation, a random error was injected into the deterministic model; which is the system error $\mathbf{p}_t$. Since this error or noise arises from process finite difference operator **A** (state transition matrix), called the process noise. This can be estimated by the difference between the model prediction and the optimal estimate of the true value. The system error $\mathbf{p}_t$ is assumed to have covariance matrix **Q**. Since the transport process occurs in a geographically specific area, $\mathbf{p}_t$ is a function of the hydrogeology specific to the area where the transport of the pollutant takes place, meaning that the system error must be correlated regionally. Without the true value, the system error cannot be estimated. However, the probabilistic nature of the error can be estimated by repetitive model calibration. For example, one can perform the pollutant transport experiment in a particular field by varying the flow parameters and initial inputs. With those results, the deviation of experimental results can be predicted easily from deterministic model results which indicate the system error $\mathbf{p}_t$. By calculating the standard deviation of system errors, the system error covariance matrix **Q** (for a geologically specific field) can be easily obtained. This kind of calibration is called off-line sampling. Conwell et al. (1997), Chang and Jin (2005), and Webster and Oliver (1992) used a time-independent Gaussian system error with $\sigma_{sys} = 8$ mg/L. In this study, $\sigma_{sys} = 10$ mg/L is used as the standard deviation of the system error. Since our one-dimensional transport model deals with only one horizontal point, there is no practical scope to implement regional noise in the experimental space domain. The Gaussian distribution or the normal distribution is the most widely used family of distributions in statistics and many statistical tests are based on the assumption of normality. When the system is highly nonlinear, there is greater tendency for the system to be non-Gaussian. Any contaminant transport system with an instantaneous pollutant source provides an exponential concentration profile. However, our system is not highly nonlinear. Therefore, it is more reasonable to assume the noise as Gaussian than other distributions. Consequently, a time independent Gaussian system error with $\sigma_{sys} = 10$ mg/L is injected into the deterministic model to construct the system model.

The process or system equation can be expressed as,

$$\mathbf{x}_{t+1} = \mathbf{A}_t \mathbf{x}_t + \mathbf{p}_t, \quad t = 0,1,2,3,4,\ldots\ldots m \qquad (5)$$

Here, m = 50 and $\mathbf{p}_t$ are the model system error and process noise, respectively. The model error $\mathbf{p}_t$ is taken from a Gaussian distribution which has a zero mean and standard deviation of 10 mg/L. The expected value of error is zero. $E\{\mathbf{p}_t\} = 0$ and $E\{\mathbf{p}_t \mathbf{p}_l^T\} = \mathbf{Q}\delta_{tl}$. The subscript of error vector **p** denotes the time index. Here δ_{tl} is a Dirac's delta function having a value either 0 or 1. $\delta_{tl} = 0 \; if \; t \neq l \; and \; \delta_{tl} = 0 \; if \; t = l$. For our case, $t = l$ which gives $E\{\mathbf{p}_t \mathbf{p}_t^T\} = \mathbf{Q}$. Therefore, **Q**

is a positive-definite matrix. For our case, it is a $n \times n$ diagonal matrix, and n is the number of nodes in space, where n = 10. $\mathbf{Q} = diag[\sigma_i^2]_{n \times n}$; σ_i is the standard deviation of system noise at node i and $i = 1,2,...,10$. We assume $\sigma_i = \sigma_{sys} = 10$ for our model.

To assimilate the actual field information optimally by filtering approach, observation data is required. Field measurements (observations) are usually quite limited and may be used only to calibrate numerical models and to estimate different flow parameters. Observations in the field may not be the true estimate of the state. For example, two types of observed error can be included here. One is instrumental error; different instruments serving the same purpose might provide different results at the same time. Again different person can measure different value with the same instrument. Therefore, the observation is affected with noise or error which occurs randomly. Observational error can be obtained from the analysis of historical data and measurement calibration. To simulate this phenomenon, we injected a random error with Gaussian distribution having a standard deviation of 5 mg/L. This is defined as an observation error or observation noise, and it can be also termed as "measurement noise". Pinder (1973) used observations in a ground water modeling study of a plume of dissolved Chromium in a sand and gravel aquifer. A computer model was used to yield a reasonable reproduction of the observed contaminant plume. The solution series was termed as "measurement data". Chang and Jin (2005) used an assumed true value and a numerical random scheme to create the "measurement data" in subsurface contaminant transport modeling; as well regional measurement noise where the highest measurement noise covariance element value was 9 mg/L. Observation represents the true field data, if there is no measurement or observation error. In other words, observation must represent the true state with a certain amount of accuracy depending on measurement accuracy. To undertake the simulation process and to be logistically acceptable, an analytical solution of the governing PDE is taken as the true value of the contaminant state. The measurement noise is added to the analytical result to make observation or measurement models. Although in the field, we usually have fewer observation locations with respect to all simulation grid points. For this study, simulated observations are used at each of the study nodes. Therefore, the observation data pattern matrix **H** becomes a 10 × 10 identity matrix. For this model, only one observation point in space is adequate. The observation or measurement equation can be expressed as,

$$\mathbf{z_t} = \mathbf{H}\mathbf{x_t^T} + \mathbf{O_t}, \quad t = 0,1,2,3,4,..... \; m \qquad (6)$$

Observation $\mathbf{z_t}$ is a vector having the observed state values of all nodes at time step t. The superscript T over state vector $\mathbf{x_t}$, denotes the true value of the state. Observation $\mathbf{z_t}$ can be simulated by data pattern matrix **H** and observation error $\mathbf{O_t}$. If we have the same data at all the nodes we are dealing with, **H** will be an identity matrix having $n \times n$ dimension, where n is the total number of nodes in the model domain. The error vector $\mathbf{O_t}$ is assumed to be white noise having a covariance matrix $\mathbf{R_t}$. The observation error vector $\mathbf{O_t}$ is constructed by a vector having the elements from a Gaussian distribution with a zero mean and a standard deviation of 5 mg/L. **R** is a positive-definite matrix. For our case, it is $n \times n$ diagonal matrix, and n is the number of nodes in space where, n=10. $\mathbf{R} = diag[\zeta_i^2]_{n \times n}$; ζ_i is the standard deviation of observation noise at node i and $i = 1,2,...,10$.

The determination of the process noise covariance **Q** is usually very difficult as typically the ability to directly observe the process estimated is absent. Sometimes, a relatively simple (poor) process model can produce acceptable results if one "injects" enough uncertainty into the process via the selection of **Q** (Welch and Bishop, 1995). Process noise covariance **Q** is the function of the hydrogeology of the field where the pollutant transport takes place. The deviation of the deterministic model from the true state depends on the accuracy of the parameters used and the heterogeneity of the subsurface. If the soil layer is uniform and the flow parameters remain constant during the transport, the system error $\mathbf{p_t}$ will be close to zero. Then the process noise covariance **Q** will become negligible. However, the real field **Q** must have some numerical values other than zero due to the random behavior of the field geology and flow parameters. To implement filtering, the statistical structure of the process noise must be determined for the specific field where the contaminant transport takes place. An experiment can be designed to measure **Q** and provide a particular matrix. Since it is beyond our capacity to predict the process noise covariance **Q** experimentally, a spatially independent Gaussian process noise with 10 mg/L standard deviation is selected to undertake the simulation process. Again, **Q** is sometimes changes dynamically during filter operation in order to adjust to different dynamics. However, it is a common

practice to assume that **Q** is constant in dynamic state estimation problems (Welch and Bishop, 1995). In this study, **Q** is kept constant throughout the filtering operation.

The measurement noise covariance $\mathbf{R}_t$ is generally possible and practical to calculate. The measurement noise covariance can be predicted by examining the performance of the measuring device and measuring procedures. On-line state estimation is done by applying two filtering approaches. It is usual to take some off-line sample measurements in order to determine the variance of the measurement noise while operating the filter. In this study, measurement noise $\mathbf{O}_t$ is taken from a Gaussian distribution with 5 mg/L standard deviation. The corresponding noise covariance matrix is constructed before the operation of filters. The measurement noise covariance $\mathbf{R}_t$ is kept constant throughout the filtering operation.

Data Assimilation with Kalman Filtering

Consider a true subsurface state variable x. A KF scheme considers the estimate x^E to be a linear combination of predicted value x^P and the observed value x^O.

$$x^E = k_P x^P + k_O x^O \tag{7}$$

Here k_P and k_O are weighing coefficients for the x^P and x^O respectively. The estimate x^E is considered to be optimal if the mean square error (MSE) of x^E is minimum.

$$MSE = E(x^E - x)^2 \tag{8}$$

where E is the expectation value operator. If the assumption of an unbiased estimate requirement is satisfied, then

$$k_P = \frac{MSE_O}{(MSE_P + MSE_O)} \tag{9}$$

$$k_O = \frac{MSE_P}{(MSE_P + MSE_O)} \tag{10}$$

The optimal estimate according to KF will then be

$$x^E = x^P + k_O(x^O - x^P) \tag{11}$$

Now consider and recall the process equation and observation equation respectively,

$$\mathbf{x_{t+1}} = \mathbf{A_t x_t} + \mathbf{p_t},\ t = 0,1,2,3,4,\ldots m$$

$$\mathbf{z_t} = \mathbf{H x_t^T} + \mathbf{O_t}, t = 0,1,2,3,4,\ldots m$$

Using the basic idea of equation (7), the optimal estimator by KF is

$$\mathbf{x_{t+1}(+)} = \mathbf{x_{t+1}(-)} + \mathbf{k_{t+1}(z_{t+1} - H x_{t+1}(-))} \tag{12}$$

where, $\mathbf{x_{t+1}(+)}$ is the estimated value after the KF adjustment, and $\mathbf{x_{t+1}(-)}$ is the state before the KF adjustment, i.e. the predicted value from the model. The matrix $\mathbf{k_{t+1}}$ is determined by

$$\mathbf{k_{t+1}} = \mathbf{P_{t+1}(-)H^T} (\mathbf{HP_{t+1}(-)H^T} + \mathbf{R_{t+1}})^{-1} \tag{13}$$

where $\mathbf{P_{t+1}}$ is the optimal estimate error covariance matrix and can be calculated by

$$\mathbf{P_{t+1}(+)} = \mathbf{P_{t+1}(-)} - \mathbf{P_{t+1}(-)H^T} (\mathbf{HP_{t+1}(-)H^T} + \mathbf{R})^{-1} \mathbf{HP_{t+1}(-)} \tag{14}$$

$$\mathbf{P_{t+1}(-)} = \mathbf{AP_t(+)A^T} + \mathbf{Q_t} \tag{15}$$

Here $\mathbf{K_{t+1}}$ called the Kalman optimal gain or Kalman filter determines how much the estimated value can gain using this filtering system from the observations.

Equation (5), (6) and equation (12) to (15) are the six basic equations of KF. To predict the optimal state by KF, $\mathbf{x_{t+1}}$ of equation (5) is used and $\mathbf{x_t}$ is substituted by $\mathbf{x_{t+1}}$ in equation (6) to get $\mathbf{z_{t+1}}$. Then using Equation (15), (14), (13), and (12) sequentially, the optimal estimator $\mathbf{x_{t+1}(+)}$ is estimated. This value of $\mathbf{x_{t+1}}$ will be used to predict next time step state of **x** (i.e. $\mathbf{x_{t+2}}$) by means of the above equations. This recursive operation will continue up to the expected time step.

Data Assimilation by PF with SIR (Sequential Importance Re-sampling)

Particle filters are sequential Monte Carlo methods that estimate system states from a state space model given the measurement sequence. The probability distribution function of the system state can be inferred. The basic idea of PF is to approximate the distribution, $p(x_t \mid z_{1:t})$ using a set of random samples called particles. Let $z_{1:t} = (z_i,\ i = 1, 2, \ldots t)$ be the measurement sequence and assume that the prior distribution $p(x_o)$ is known, the posterior probability can be obtained sequentially by prediction and update. The prediction and update equations are as follows:

$$p(x_t \mid z_{1:t-1}) = \int p(x_t \mid x_{t-1}) p(x_{t-1} \mid z_{1:t-1}) dx_{t-1} \tag{16}$$

$$p(x_t \mid z_{1:t}) = \frac{p(z_{1:t} \mid x_t) p(x_t \mid x_{t-1})}{p(z_t \mid z_{1:t-1})} \tag{17}$$

The above equations are the optimal solutions from a Bayesian perspective to the non-linear state estimation problem. One limitation is that the evolution of the posterior density generally cannot be determined analytically. Thus, some approximations must be made. PF approximates the posterior probability distribution, $p(x_t \mid z_{1:t})$ by a set of supporting random samples (particles) $x_t^i, i = 1, 2, \ldots\ldots, N$, with associated weights

$$w_t^i : p(x_t \mid z_{1:t}) \approx \sum_{i=1}^{N} w_t^i \delta(x_t - x_t^i) \quad (18)$$

where $\delta(x)$ is an indicator function known as Dirac's delta function which is equal to unity if x = 0 and otherwise equal to zero. The weights sum to zero. The filtered state is taken as the mean of posterior density. The next issue is how to determine the weights of the particles. Since $p(x_t \mid z_{1:t})$ is not in the form of a traditional probability density function, weights cannot be assigned by direct sampling, but determined using importance sampling. An importance density $q(x_k \mid Z_{1:k})$ is defined from which samples drawn. Thus the weights are defined as:

$$w_t^i \propto \frac{p(x_t^i \mid z_{1:t})}{q(x_t^i \mid z_{1:t})} \quad (19)$$

If the importance density is selected appropriately and only dependent on the current observation, z_t and the past state, x_{t-1} the weights can be updated as follows(Arulampalam et al., 2002):

$$w_t^i \propto w_{t-1}^i \frac{p(z_t \mid x_t^i) p(x_t^i \mid x_{t-1}^i)}{q(x_t^i \mid x_{t-1}^i, z_t)} \quad (20)$$

With these particles and associated weights, the optimal state can be estimated by a normalized summation. The mean of the states can be approximated by $\overline{x_t} = \sum_{i=1}^{N} w_t^i x_t^i$. To implement PF, generally two implementation issues are considered (Chen et al., 2004); the first of which is degeneracy and the other is how to choose importance density. After some iterations degeneracy occurs, when only one particle has significant weight and all other particles weigh zero. Thus considerable computational effort will have been spent on updating particles whose contribution to the approximation of $p(x_t \mid z_{1:t})$ is negligible. Re-sampling can be used to eliminate those particles with small weights thereby focusing on particles with large weights. Re-sampling generates set $\langle x_k^i, i = 1 \ldots\ldots\ldots N \rangle$ with Pr $(x_t^j = x_t^i) = w_t^i$. Here j is the particle index after resampling. The parent relationship is denoted, parent $(j) = i$. The weights are reset to be 1/N as the samples are independent and identically distributed and then drawn from a discrete density function. By re-sampling, those particles with zero weight will be discarded. $q(x_t^i \mid x_{t-1}^i, z_t)$ is used as the importance density. Using $q(x_t^i \mid x_{t-1}^i, z_t) = p(x_t^i \mid x_{t-1}^i)$ yields a simple form to update the weights according to equation (20): $w_t^i \propto w_{t-1}^i (z_t \mid x_t^i)$. A PF with this importance density and re-sampling step is called a sequential importance re-sampling (SIR) filter. A standard SIR PF uses the prior distribution as the importance density.

Figure 2 describes the methodology followed in this simulation process. The mathematical mechanistic model of contaminant transport is represented by the partial differential equation expressed in equation (1)which is discretized by the FTCS method to get difference equations. These equations can be regarded as general state space models with the difference variables defining the states. The state space model represented in equation (4) is incorporated with independent and identically distributed process noise to get the process or system equation. The analytical solution of the transport model is incorporated with independent and identically distributed measurement or observation noise to simulate the observation equation. To run the filtering four inputs are needed: process equation, process noise variance, observation, and measurement noise variance. The statistical nature of process noise and observation noise for our model is described earlier. With those four inputs, both KF and PF techniques are implemented to get the optimal state.

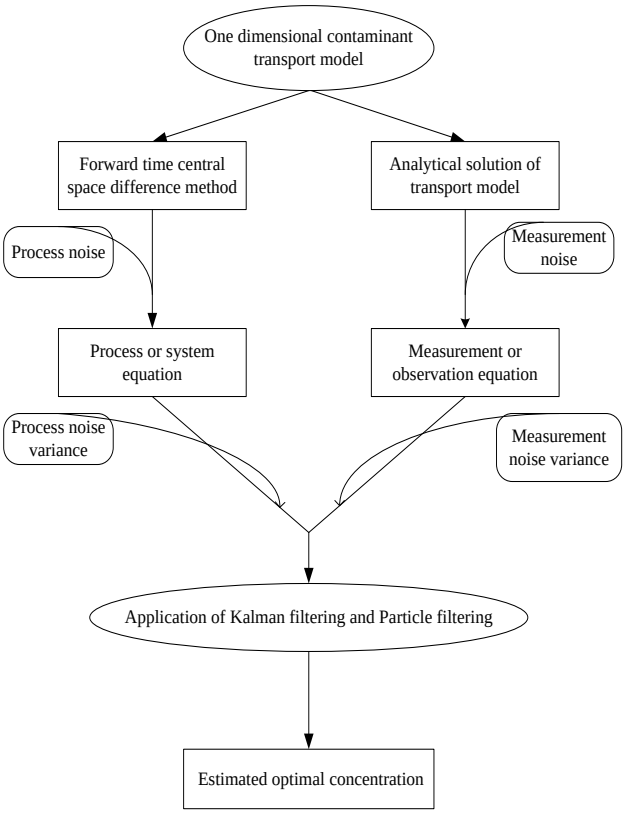

FIG. 2 AN OVERVIEW OF THE PROCEDURE FOLLOWED IN THE STUDY

Illustration Example

A desktop study is undertaken to simulate a synthetic industrial leachate transport field. In this study, one-dimensional leaching of benzene from an industrial landfill site is simulated by KF and PF data assimilation scheme. Figure 3 shows that vertically 10 nodes with 0.02 m spacing are taken in model space domain. Each time interval is taken as 0.01 day. Those space and time interval are chosen after stability analysis of the numerical solution of the problem.

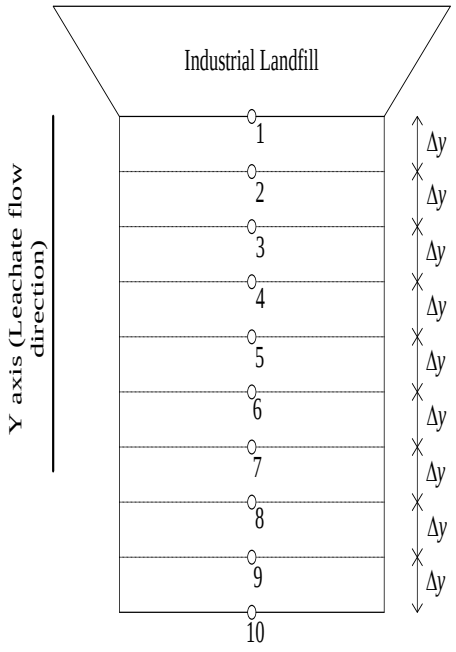

FIG. 3 SCHEMATIC DIAGRAM OF BENZENE LEACHATE FLOW SHOWING 10 DISCRETE NODES

Here λ is called diffusion number. In this model, the benzene contaminant mass at the source, M_o, is 16.5 gm $y = 0$ and $t = 0$ in a column with a cross-sectional area of 1 m^2. The longitudinal dispersivity, α, is 0.01 meters. A water flux of 0.2 m^3/d is maintained continuously at $y = 0$. The porosity, η, of the leaching soil medium is 0.25, which gives tortuosity of porous medium, $\Gamma = 0.63$. Flow velocity, V, is calculated as 0.8 m/day vertically downward. Molecular diffusion, D_m = 1.44E-06 m^2/day, gives the dispersion of the contaminant, D_y = 0.008001 m^2/day. Fraction organic Carbon = 0.0001, soil solid specific gravity = 2.65, organic carbon/water partition coefficient = 83 L/Kg, bulk density = 1.99 g/cm^3, and retardation factor = 1.1 are used in the study. The retardation factor (R) for benzene leaching into the specified soil medium is calculated by using an EPA on-line tool for site assessment. For industrial landfill leachate, typical concentration of benzene is about 2000 mg/L and maximum concentration may reach 39000 mg/L (Tchobanoglous et al., 1993). Initial concentration (C_o) is calculated as about 1985 mg/L for pulse or instantaneous source which is injected at the top node (i.e. node no. 1). If we deal with n discrete nodes in the space, the dimension of the state vector will be $n \times 1$. Here $n = 10$. Therefore, the initial state vector x_o is a 10×1 vector with C_o in the first cell and zero in the nine other cells.

The Filtering Effectiveness Examination

The effectiveness of filtering and deterministic model is measured by comparing the model predicted results with the true value. Since all noise distributions are assumed to be normally distributed, the root mean square error (RMSE) is used as the effectiveness parameter.

$$RMSE(t) = \sqrt{\frac{1}{n-1}\sum\left[C_E(i,t) - C^T(i,t)\right]^2} \quad (21)$$

where $RMSE\ (t)$ = error (mg/L) at time step t; $C_E\ (i, t)$ = expected value of concentration at node i at time step t; $C^T\ (i, t)$ = observed value of concentration at node i at time step t; n = total number of nodes.

Results and Discussions

At the first step of the experiment, the concentration profile is constructed without observation correction. To deal with the true field, the state transition matrix, $\mathbf{A_t}$, should be different for each time step. Since all parameters used in this experimental scheme are kept constant with time, the state transition matrix does not change with time. That means $\mathbf{A_t} = \mathbf{A_{t-1}} = \ldots \ldots = \mathbf{A_1} = \mathbf{A}$. The state transition matrix changes the state from one time step to the next. The experimental scheme is designed to use 10 discrete nodes with Δy spacing in the space domain and 50 time steps with Δt magnitude in the space domain. In this way, the contaminant state is estimated by constructing 50 state vectors. Along the timeline, the peak in the concentration profile moves along from the first node to the lower nodes. This indicates that the velocity direction of pollutant transport is from the top layer of soil to the lower layer meaning the transport direction along the y axis as depicted in Figure 3. Again, the concentration profiles give bell-shaped curves (Figure 4 and 5), indicating the instantaneous or pulse input of the contaminant.

In our model, no initial error of the state is injected. The same initial concentration is provided for the deterministic (FTCS) model, KF, and PF. Therefore, the initial optimal error covariance matrix ($\mathbf{P_o}$) is a zero

matrix. Under these conditions where **Q** and **R** remain constant throughout the KF operation, both the estimation error covariance $\mathbf{P}_t$ and the Kalman gain $\mathbf{K}_t$, quickly stabilizes. The KF algorithm is processed with its basic six equations as described in the methodology. The KF requires six inputs at the beginning of the operation of the filter, namely, initial state vector $\mathbf{x}_0$, process noise $\mathbf{P}_t$, measurement noise $\mathbf{O}_t$, process noise covariance **Q**, measurement noise covariance **R**, and state transition matrix **A**. The probabilistic analysis has been done by recursive calculation of its six equations which are in the matrix form. For data assimilation by PF, 500 random samples are drawn from the prior density function in each time step. Then, by using the observation model it develops a posterior density function of the state. Finally, the mean value of the probability distribution gives the optimal state value. The basic inputs of PF are as same as those of KF. First, the deterministic model was run for verification and comparison purposes. The profile given is a smooth curve which represents the theoretical deterministic behavior of the solution process. The experimental scheme was designed in such a way that, at each time step, both KF and PF combine the system model and observation model to simultaneously provide the next prediction. In contrast to the smooth curve of the FTCS result, KF and PF give concentration profiles with irregular shape which represents more realistic field behavior. Due to the randomness and uncertainties of the true field, its concentration profile generally gives oscillated curves rather than smooth ones.

The comparison between the deterministic model and the two filtering model results is shown in Figure 4 to 7. The experiment shows that the benzene concentration is substantially reduced in the study space domain at higher time steps. It is difficult to compare different methods at lower levels of concentration. Therefore, comparison is done by taking up to 25 timestep results. The comparison is constructed on the basis of fixed space and fixed time criteria. Figure 4 and 5 show the benzene concentration profiles at fixed locations while Figure 6 and 7 show the profiles at fixed time steps. Again the model was run with and without numerical dispersion corrections. Figure 4 shows that the true peak concentration at node 3 is about 900 mg/L and that of the FTCS method is about 670 mg/L. Therefore, the peak concentration difference between the true value and the FTCS method is about 230 mg/L in the case without any numerical dispersion correction. For FTCS, the peak is located at almost the same time step as the true value. On the other hand, Figure 5 shows that the peak concentration for FTCS is almost the same as the true value in the case with numerical dispersion correction. The difference between two values is about 50 mg/L. But the peak occurs with a time-lag of 3 time steps. The prediction resulting from filtering, which is corrected by observation data, seems closer to the true value than that from the deterministic FTCS model. It is very difficult to differentiate which filtering method, KF or PF, performs better at this fixed node. But both filtering results fit well with the true value whereas the FTCS results deviate more from the true state than that of filtering. Figure 6 and 7 depict the comparison of the deterministic and filtering models with the true state after a 0.15 day with and without numerical dispersion corrections respectively. Figure 6 shows that the peak value of the benzene concentration profile takes place at the 6th node and the difference between the true value and FTCS method is about 100 mg/L. On the other hand, Figure 7 shows that the peak value takes place at the 5th node in the case of the FTCS method; the difference between the true value and FTCS method is more than 100 mg/L. Thus, the model with dispersion correction shows more discrepancy between true value and FTCS value than the model without dispersion correction. Again, like the experimental results at the fixed distance, the simulation results from filtering matches well with the true field results at the same timestep. The deterministic model provides more deviated results from the true field than that of filtering. Thus, filtering outperforms the deterministic model quantitatively.

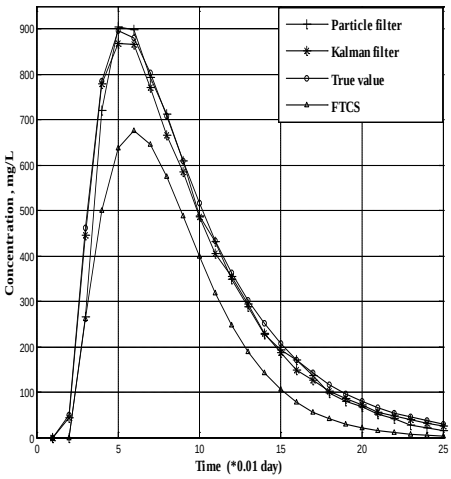

FIG. 4 BENZENE CONCENTRATION PROFILE AT NODE 3 WITHOUT NUMERICAL DISPERSION CORRECTION

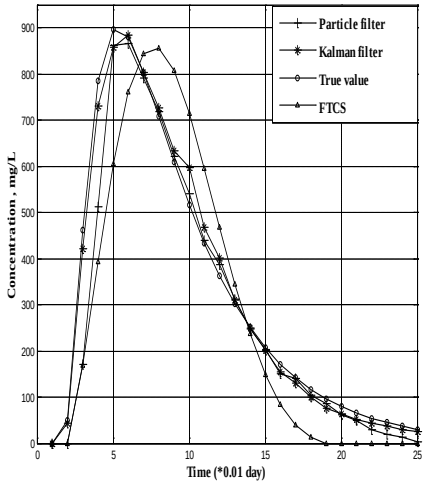

FIG. 5 BENZENE CONCENTRATION PROFILE AT NODE 3 WITH NUMERICAL DISPERSION CORRECTION

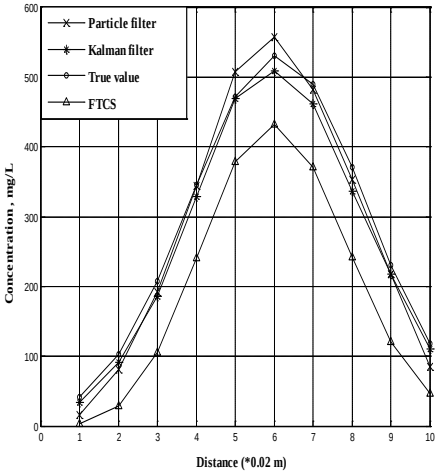

FIG. 6 BENZENE CONCENTRATION PROFILE AFTER 0.15 DAY WITHOUT NUMERICAL DISPERSION CORRECTION

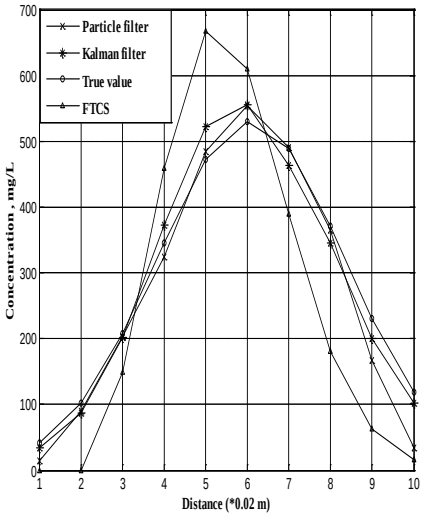

FIG. 7 BENZENE CONCENTRATION PROFILE AFTER 0.15 DAY WITH NUMERICAL DISPERSION CORRECTION

Actually, the system model is affected by numerical dispersion correction, however, the observation model is not. Therefore, the observation model remains almost the same in both cases, with and without numerical dispersion correction. In addition, the filtering procedure provides more weight on observation than the system. For this reason, the filtering results are closer to observation, which reveals the true state accordingly.

Since the model without numerical dispersion correction outperforms that with that correction, the former model was chosen to examine the filtering effectiveness by RMSE (root mean square error) evaluation. As shown in Figure 8, RMSE value for the PF prediction gradually decreases with the increase of time steps. The RMSE vs. time step graph for the PF prediction takes a hyperbolic shape which expresses a good merging tendency toward the true value. Such a shape signifies the successful data assimilation. The error starts with 135 mg/L at time step 1 and deceases to 10 mg/L at time step 22. In this way, PF prediction reduces the error from 30% to 80% from time step 1 to time step 22, respectively. On the other hand, KF provides consistent performance about RMSE. RMSE for KF prediction varies from 40 mg/L to 4 mg/L. KF reduces this error from 70% to 90%. In case of KF, that means the range of error reduction varies within a 20% range whereas error reduction of PF varies within a 50% range. For the first several time steps, the prediction error of KF is 30% to 50% less than that of PF.

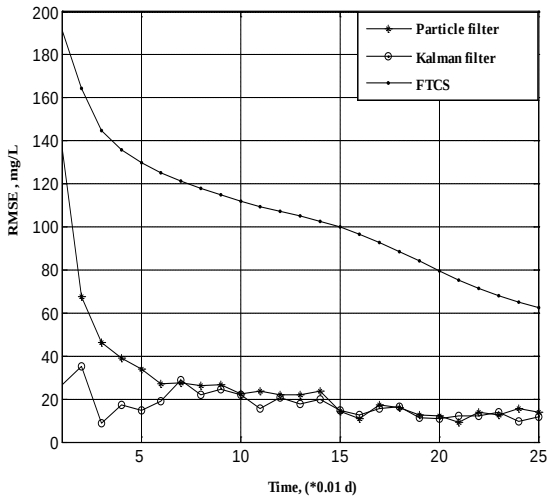

FIG. 8 PREDICTION ERROR (ROOT-MEAN-SQUARE ERROR) FOR DATA ASSIMILATION WITH DETERMINISTIC MODEL (FTCS), KF AND PF

Conclusion

The results of this study show that estimation

techniques using KF and PF offer a more accurate solution than conventional numerical approach for model development with noisy and incomplete data. Although 2-D and 3-D models become more complicated in mechanism expression and in computational implementation, the conclusions obtained from this simple one dimensional model can be applied to the models with higher dimensions, because mathematical systems remain the same.

The true value of the state takes place in a space for only one time due to the randomness of the field behavior. It is impossible to relate all hydro geological parameters mathematically in a deterministic model so that it can give the true solution. Therefore, instead of the true value, the best engineering approach in contaminant transport modeling is to get the solution as close to the true value. Since a true estimate of the contaminant state cannot be achieved, any possible estimate that is the closest to the true estimate can be treated as a better solution for the contaminant transport system. For benzene transport, both KF and PF can provide accurate results for the illustrated case. In this study, the contaminant transport problem is treated as a dynamic state estimation problem. The Bayesian approach, like KF and PF, is based on the probabilistic state space formulation and requires update of information on receipt of new measurements (observations). The traditional deterministic model does not require either probabilistic formulation of the state or the update of the state with new measurements. Therefore, the program algorithm becomes more complex in the case of filtering approach than that of the conventional deterministic model. As a result, filtering requires more computational effort compared to the conventional numerical solution.

The experiment is designed and constructed in such a way that every time the program is run, different random numbers are generated to simulate the random behavior of the subsurface transport plume. The RMSE prediction curve does not vary significantly and the qualitative shape remains the same for different run. The RMSE prediction is evaluated with one time run. At time step 5, the estimated RMSE of the numerical solution is 7.5 times greater than that of KF and 3.5 times greater than that of PF. It is indicated that KF and PF outperform the numerical solution. The discrete KF has six equations that sequentially provide the recursive estimation of the state in a matrix form. This provides a rigorous general framework for the dynamic state estimation problem with linear systems and Gaussian distribution. On the other hand, particle filters are appropriate to handle general dynamic state space models and do not rely on the assumptions of linearity or a Gaussian posterior density. However, PF needs the probabilistic formulation of the state. The problem algorithm and formulation is more complex in case of PF than that of KF. Therefore, PF requires more computational efforts than that of KF. For this one dimensional case, KF outperforms PF. On an average, KF can reduce the error from 70% to 90% and PF can reduce the error from 30% to 80% as compared to the conventional numerical approach.

ACKNOWLEDGMENTS

This work was sponsored by the Department of Energy Samuel Massie Chair of Excellence Program under Grant No. DE-NA0000718. The views and conclusions contained herein are those of the writers and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the funding agency.

Notation

The following symbols are used in this paper:

A = cross-sectional area; $\mathbf{A}_t$ = state transition matrix at time step t; C = solute concentration, mgl^{-3}; $C_E\ (i,\ t)$ = expected value of concentration at node i at time step t; $C^T\ (i,\ t)$ = observed value of concentration at node i at time step t; $\mathbf{C}_{i,t}$= vector of pollutant concentration at node i and time t; D_y = dispersion coefficients in y direction, m^2s^{-1}; E = expectation value operator; $\mathbf{H}$ = observation data pattern matrix; $\mathbf{K}_{t+1}$ = Kalman optimal gain or Kalman filter at time step $t+1$; MSE = mean square error; N = total number of random samples, (here N = 500); n = total number of nodes; $\mathbf{O}_t$ = observation error at time step t; $\mathbf{p}_t$ = model system error or process noise; $\boldsymbol{P}_{t+1}$ = optimal estimate error covariance matrix at time step $t+1$; $p(x_t \mid z_{1:t})$ = conditional probability of x_t given the observation sequence $z_{1:t} = (z_i,\ i = 1, 2, \ldots, t)$; $\mathbf{Q}_t$ = model system error covariance matrix at time step t; R = retardation factor, dimensionless; $RMSE$ = Error (mg/L) at time step t; $\mathbf{R}_t$ = observation noise covariance at time step t; t = time, s; V = linear velocity of flow field in the y direction, ms^{-1}; w^i_t = associated weight of x^i_t; $\mathbf{x}_t$= the state variable described as a vector of contaminant concentration of all nodes at time t; x^i_t = supporting random samples (particles) regarding to each element of state vector $\mathbf{x}_t$, i =1, 2, … , N; x^E = estimated state variable by KF; x^O = observed value; x^P = predicted value of state vector; $\overline{x_t}$ = optimal mean state after

normalized summation; y = cartesian coordinate direction along the flow line, m; $\mathbf{z}_t$ = vector having the observed values of state of all nodes at time step t; η = porosity of soil medium; σ_i = standard deviation of process noise at node i; σ_{sys} =standard deviation of the system error; $\delta(x)$ =Dirac's delta function; ζ_i = standard deviation of observation noise at node i.

REFERENCES

Arulampalam, S., S. Maskell, N.Gordon, and T. Clapp. "A tutorial on Particle Filters for on-line non-linear/ non-gaussian Bayesian Tracking." *IEEE Transactions on Signal Processing* 50(2002):174-188.

Chang, S.Y., and A. Jin. "Kalman Filter with regional noise to improve accuracy of contaminant transport models." *Journal of Environmental Engineering* 131(2005): 971-982.

Chen, T., J. Morris, and E. Martin. "Particle Filters for the estimation of a state space model." Proc. European Symposium on Computer-Aided Process Engineering 14. Computer-Aided Chemical Engineering, Lisbon, Portugal, 18(2004), 613-618.

Chen, T., J. Morris, and E. Martin. "Particle Filters for state and parameter estimation in batch processes." *Journal of Process Control* 15(2005): 665-673.

Cheng, X. "Kalman Filter scheme for three- dimensional subsurface transport simulation with a continuous input." MS thesis, North Carolina A&T State University, Greensboro, NC, 2000.

Conwell, P. M., S. E. Silliman, and L. Zheng. "Design of a piezometer network for estimation of the variogram of the hydraulic gradient: The role of the instrument." *Water Resource Research* 33(11)(1997): 2489-2492.

Cosby, B. J., G. M. Hornberger, and M. G. Kelly. "Identification of photosynthesis-light models for aquatic systems II: Application to a macrophyte dominated stream." *Ecological Modeling* 23(1984): 25-51.

Encyclopedia. "Benzene: Encyclopedia." Last modified August 27, 2009. http://en.allexperts.com/e/b/be/ benzene. html.

Ferraresi, M., and A. Marinelli. "An extended formulation of the integrated finite difference method for ground water flow and transport." *Journal of Hydrology* 175 (1996): 453-471.

Graham, W., and D. McLaughlin. "Stochastic analysis of non-stationary subsurface solute transport: 2. Conditional moments." *Water Resources Research* 25(11) (1989), 2331-2355.

Harrouni, K.EI, D. Ouazar, L.C. Wrobel, and A. H. D. Cheng. "Aquifer parameter estimation by extended Kalman Filter and boundary elements." *Engineering Analysis with boundary Elements* 19(3)(1997):,231-237.

Li, L., H. Zhou, H.-J. Hendricks Franssen, and J. J. Gómez-Hernández. "Modeling transient groundwater flow by coupling ensemble Kalman filtering and upscaling." *Water Resources Research* 48(1)(2012): W01537.

McLaughlin, D. "An integrated approach to hydrologic data assimilation: interpolation, smoothing andforecasting." *Advances in Water Resources* 25(2002): 1275-1286.

Panzeri, M., M. Riva, A. Guadagnini, and S. P. Neuman. "Data assimilation and parameter estimation via ensemble Kalman filter coupled with stochastic moment equations of transient groundwater flow." *Water Resources Research* 49(3)(2013): 1334-1344.

Pinder, G. "A Galerkin-finite element simulation of Groundwater contamination on Long Island, New York." *Water Resources Research* 9(1973): 1957-1669.

Porter, D., G. Bruce, W. Jones, P. Huyakorn, L. Hamm, and G. Flach. "Data fusion modeling for groundwater systems." *Journal of Contaminant Hydrology* 43(2000): 303-335.

Rozos, E., and D. Koutsoyiannis. "Benefits from using Kalman filter in forward and inverse groundwater modelling." Proc., European Geosciences Union General Assembly, Geophysical Research Abstracts, Vienna, 13 (2011): 2011-2212.Schrader, B. P., and S. F. Moore. "Kalman filtering in water quality modeling: theory vs. Practice." Proc., 9th conference on Winter Simulation, Gaitersburg, Maryland, 2 (1977): 504-510.

Schwartz, F. W., and H. Zhang. *Fundamentals of groundwater*. New York: John Wiley & Sons, Inc., 1994.

Tchobanoglous, G., H. Theisen, and S. A. Vigil. *Integrated solid waste management*. New York: McGraw-Hill, 1993.

Van Geer, F.C. (1982). "An equation based theoretical approach to network design for ground water levels using Kalman Filter." *International Association of Hydrological Science* 136(1982): 2411-250.

Walker, J. P., G. R. Willgoose, and J. D. Kalma. "One-dimensional soil moisture profile retrieval by assimilation of near-surface observations: a comparison

of retrieval algorithms." *Advances in Water Resources* 24(6) (2001), 631-650.

Webster, R., and M. A. Oliver. "Sample adequately to estimate variograms for soil properties." *Journal of Soil Science* 43 (1992), 177-192.

Welch, G., and G. Bishop, (1995). "An introduction to the Kalman Filter." University of North Carolina, Department of Computer Science, TR 95-041 (1995). Last accessed March 10, 2009. http://www.cs.unc.edu/ ~tracker /ref/s2001.

Whitehead, P. G., and G. M. Hornberger. "Modeling Algal behavior in the river Thames." *Water Resource Research* 18(8) (1982): 945-953.

Yangxiao, Z., C. B. M. Te Stroet, and F. C. Van Geer. "Using Kalman Filter to improve and quantify the uncertainty of numerical ground water simulation: application to monitoring network design." *Water Resource Research* 178(1991): 1995-2006.

Yu, Y.S., M. Heidari, and W. Guang- Te. "Optimal estimation of contaminant transport in ground water." *Water Resources Bulletin* 25 (1989): 295-300.

Zou, S., and A. Parr. "Optimal estimation of two-dimensional contaminant transport." *Ground Water* 33(2) (1995): 319-325.